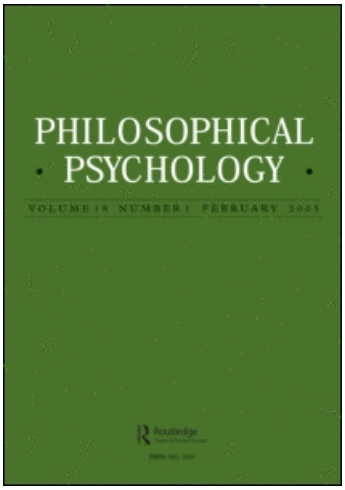
Neural models that convince: Model hierarchies and other strategies to bridge the gap between behavior and the brain.

Meeter, M.; Jehee, J.F.M; Murre, J.M.J.

published in
Philosophical Psychology
2007

DOI (link to publisher)
10.1080/09515080701694128

document version
Publisher's PDF, also known as Version of record



Philosophical Psychology
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713441835

Neural Models that Convince: Model Hierarchies and Other Strategies to Bridge the Gap Between Behavior and the Brain

Martijn Meeter; Janneke Jehee; Jaap Murre

To cite this Article Meeter, Martijn , Jehee, Janneke and Murre, Jaap(2007) 'Neural Models that Convince: Model Hierarchies and Other Strategies to Bridge the Gap Between Behavior and the Brain', Philosophical Psychology, 20: 6, 749 — 772
To link to this Article: DOI: 10.1080/09515080701694128
URL: http://dx.doi.org/10.1080/09515080701694128

Philosophical Psychology
Vol. 20, No. 6, December 2007, pp. 749–772

Neural Models that Convince: Model Hierarchies and Other Strategies to Bridge the Gap Between Behavior and the Brain

Martijn Meeter, Janneke Jehee and Jaap Murre

Computational modeling of the brain holds great promise as a bridge from brain to behavior. To fulfill this promise, however, it is not enough for models to be 'biologically plausible': models must be structurally accurate. Here, we analyze what this entails for so-called psychobiological models, models that address behavior as well as brain function in some detail. Structural accuracy may be supported by (1) a model's a priori *plausibility, which comes from a reliance on evidence-based assumptions, (2) fitting existing data, and (3) the derivation of new predictions. All three sources of support require modelers to be explicit about the ontology of the model, and require the existence of data constraining the modeling. For situations in which such data are only sparsely available, we suggest a new approach. If several models are constructed that together form a hierarchy of models, higher-level models can be constrained by lower-level models, and low-level models can be constrained by behavioral features of the higher-level models. Modeling the same substrate at different levels of representation, as proposed here, thus has benefits that exceed the merits of each model in the hierarchy on its own.*

Keywords: Brain, Behavior; Computational Modeling; Hippocampus; Neural Networks; Vision

1. Introduction

Most psychologists and neuroscientists agree that the brain produces behavior, and that ultimate theories of behavior will be ones that spell out the link between the two. One type of research that can function as a bridge between brain and behavior is

Correspondence to: Martijn Meeter, Vrije Universiteit Amsterdam, Dept. of Cognitive Psychology, Vd Boechorststraat 1, Amsterdam, 1081 BT Netherlands. Email: m@meeter.nl

ISSN 0951-5089 (print)/ISSN 1465-394X (online)/07/060749-24 © 2007 Taylor & Francis
DOI: 10.1080/09515080701694128

computational modeling of the brain. In the ideal case, models incorporate brain anatomy and physiology, and show us how humans or animals solve tasks. Such models can then generate predictions on both the physiological level and the behavioral level. In practice, however, a majority of computational models still focus on one or the other. In neurobiological models, behavior is often modeled—if at all—in a very abstract way, precluding behaviorally testable predictions. In psychological models, the connection to real brain processes is often so thin as to become irrelevant. This state of affairs is not surprising. A 'mindbrain' model, one that is both adequate at the biological plane and specific about behavior, must work at levels with widely different temporal and spatial scales. It must do justice to neurons that take up about 9×10^{-18} liters and spike in less than a millisecond, but also to behavioral tasks that involve the whole brain (± 1.3 liters) and take seconds to minutes or more (Anderson, 2002; Murre & Sturdy, 1995).

In recent years, however, more and more models have been proposed that are intermediate, being neither purely functional nor tied with much precision to brain anatomy (e.g., Bogacz, Brown, & Giraud-Carrier, 2001; Botvinick, Braver, Barch, Carter, & Cohen, 2001; Deco & Rolls, 2002; Gluck & Myers, 1993; Grossberg, 2001; Hasselmo, 1995a; Jensen, Idiart, & Lisman, 1996; Lengyel, Kwag, Paulsen, & Dayan, 2005; Li, 2003; Meeter & Murre, 2004; Murre, 1996; Norman & O'Reilly, 2003; Petrov, Dosher, & Lu, 2005; Polsky, Mel, & Schiller, 2004; Raffone & Wolters, 2002; Rao, Zelinsky, Hayhoe, & Ballard, 2002; Usher & Niebur, 1996; van der Velde & de Kamps, 2001). They typically take inspiration from brain anatomy to model a particular set of behavioral data, or a particular psychological function. For such models, we will use the term psychobiological models. Some of these models have generated much enthusiasm, as they hold great promise as a bridge from the brain to behavior.

Psychobiological modeling can have several goals. A model can be presented as only one of many possible instantiations of a verbal theory. The goal of the modeler is then to show that the model produces the theorized behavior, as an existence proof that the mechanisms in the verbal theory can work as proposed (e.g., Nadel, Samsonovitch, Ryan, & Moscovitch, 2000; Xing & Andersen, 2000). Sometimes, models are said to uncover computational limitations and tradeoffs, making it possible to derive general principles that are applicable to the brain's computations (McClelland, McNaughton, & O'Reilly, 1995; Ringo, Doty, Demeter, & Simard, 1994), or they make statements about the computations that a certain brain area can perform (Treves & Rolls, 1994). A goal of modeling may also be to structure data and tease apart components that underlie superficial effects in the data (e.g., modeling retention to uncover differences in forgetting rates, Meeter, Murre, & Janssen, 2005; Rubin & Wenzel, 1996). Most often, however, the aim of the modeler is to explain existing findings and predict new ones.

There are two ways in which 'explaining of findings' can be understood. A classic interpretation of theorizing in general is that a theory summarizes and systematizes observations, allowing covering laws to be extracted from the data (Hempel, 1965). This is not the view of most modelers, who usually have the pretension that their

model corresponds to something in reality. They will claim for their model what Webb (2001) has called structural accuracy. Structural accuracy refers to how well the model represents the real mechanisms underlying the target behavior. If a model is structurally accurate, it does not reproduce the target data only because it incorporates a covering law, it does so because the mechanism producing model behavior is equivalent to that producing the data in the modeled substrate. Most modelers will thus claim that their model reproduces the data in a structurally accurate way.

'Structurally accurate' is a predicate similar to 'true': one can never be 100% certain that a model is structurally accurate, but the structural accuracy of a model can be supported. For a computational model, such support can take several forms. Traditionally, models are judged on their ability to reproduce existing findings, and by the predictions derived from the model that turn out to be true. As models of behavior, psychobiological models are most often not up to standards set by formal models. For example, few biologically-inspired network models of memory have yielded quantitative predictions on memory experiments, something that had already been achieved by functional, formal models developed in the seventies and early eighties (e.g., Raaijmakers & Shiffrin, 1981). Many psychobiological models are also published long before predictions they generate are proven correct. According to traditional criteria, psychobiological models are thus not particularly successful.

There are thus two traditional sources of model support, fitting data and making predictions that turn out to be correct, and many psychobiological models do not excell on them. Such deficiencies are often said to be compensated by a third virtue, a high 'biological plausibility'. This should make the model *a priori* plausible—which is how we will refer to this third source of support for a model. This paper can be read as an analysis of the claim of biological plausibility. We will first present a framework to understand model structure, and use this to argue that for a model to be *a priori* plausible, it must not so much contain many biological details, as much as not contain assumptions that violate biological knowledge. Then we will argue that tying the model to biology also makes it more testable, but only when the modelers are specific on the ontology of their model.

There are cases in which too little data is available to constrain the model. As a solution to this problem, we will propose a new approach to modeling: that of constructing a model hierarchy. Such a hierarchy of complementary models may optimize the way data are incorporated into the model, and could bridge the gap between the seemingly unconnected levels of biology and psychology. Models in the hierarchy can also constrain one-another, and pose restrictions on modeling where behavior and physiology pose too few.

In a first part of this paper, we discuss the structure of computational models, and the three sources of support for models mentioned above. In a second part, we argue that considerations of model support suggest two strategies for building psychobiological models: being sparse and mining biology. In a third part, we discuss model hierarchies.

2. Model Structure and Model Support

Computational modeling of the brain is the attempt to build a structure that is in some way an abstraction of the brain. Characteristics of the functioning of this abstract structure can then be derived or established through simulation. Usually, such a simulation produces model behavior that can be compared to empirical data. The term 'model behavior' should not be taken to imply that it can only be compared to behavioral data. Behavior of neurons in a model can, for example, be compared to behavior of recorded neurons in a modeled brain area.

The abstract structure of a model may be thought of as a collection of assumptions that together specify a model. In an analytical, mathematical model, these assumptions are the formulas used, the representation of the data that go into the formulas, and parameter values. In neural networks, they may be the rules governing the behavior of the nodes, the subdivisions of the network, the connection schemes incorporated in the network, the number of nodes and other parameter values (see Figure 1 for examples).

Some assumptions going into the model may be supported by evidence from the brain and behavioral sciences. We will call these evidence-based assumptions. An example in Figure 1 is the assumption in the third model that inhibition in a brain area is a function both of the input to that area (feedforward inhibition), and the neural activity within the area (feedback inhibition). Assumptions may also concern unsupported ideas about brain or behavior. These assumptions, which could possibly be true of the brain, we will refer to as 'untested assumptions'. Some of these may reflect the bold new theorizing of the modeler; we will refer to these as 'hypotheses'. An example in Figure 1 is the assumption, of the top model, that memories are copied from one memory store to the next in proportion to the number of copies in the previous store. Another untested assumption could be the value of a free parameter; it could, in principle, be true that the parameter or some counterpart in the brain has the chosen value, but it has not been tested (if the value was backed up by data, it would be an evidence-based assumption).

To build a working model, it is usually not enough to only abstract away from the brain (in fact, if all mismatch with biology would refute a model, 99% of all models would have to be rejected). Assumptions have to be added that do not model anything in the brain and are not meant to be hypotheses about the brain, but that allow the model to produce behavior. For example, in the second model in Figure 1, all neurons in the hippocampus are assumed to be connected with one another (known as the assumption of 'full connectivity'). Few would call full connectivity biologically plausible, but it features in many models as an easy way to overcome the limitations of having vastly fewer neurons in the model than there are in the brain. Other examples are hard winner-take-all dynamics (assuming that at any moment only k neurons are firing), nonoverlapping input patterns (assuming that neurons cannot be part of more than one input pattern), clamping of input patterns (assuming that nothing changes to the activity in an input layer while the network

computes an output), learning through backpropagation of error, and 'empty brains' (i.e., all connections at zero or random weights) at the outset of the simulation. Such assumptions, which are in all likelihood counterfactual, we will refer to as 'heuristic assumptions' (similar use of the term is found in economics, where demonstrably false assumptions of rationality and full information are defended as heuristic devices). Untested and heuristic assumptions together will be referred to as 'unsupported hypotheses.' Figure 2 gives more examples of each type of assumption in one model of the hippocampus.

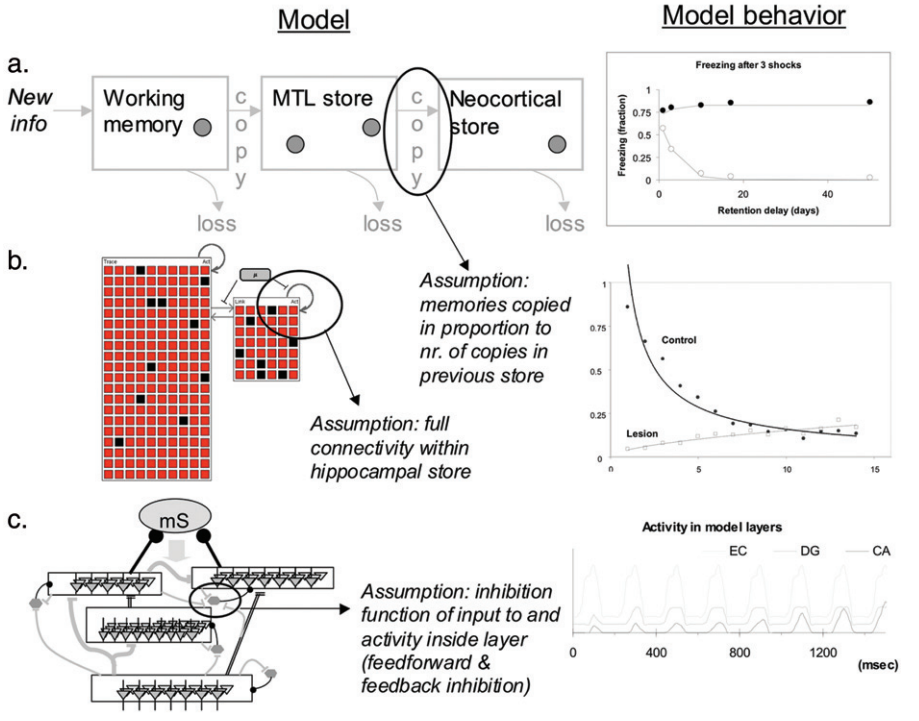


Figure 1. Examples of three psychobiological models and the behavior they produce. (a) The Memory Chain Model (Murre, Meeter, & Chessa, 2007) is an abstract model of the structure of models that can fit forgetting curves. The model structure consists of a set of chained stores, from which memories (circles) are copied to further stores or are lost. The example of model behavior is a quantitative fit of data from Frankland et al. (2001) showing rapid forgetting in mice with a gene responsible for LTP knocked out in the hippocampus. (b) The TraceLink model (Meeter & Murre, 2005; Murre, 1996) is a neural network model of corticohippocampal interactions. Model behavior consists of qualitative reproduction of patient data, shown here for the typical pattern seen in retrograde amnesia, with preferential loss of recent memories after damage to the hippocampus. (c) A low-level neural network model of the hippocampus (Meeter et al., 2004), with as typical model behavior the activity produced in different subregions of the hippocampus when a new memory is produced.

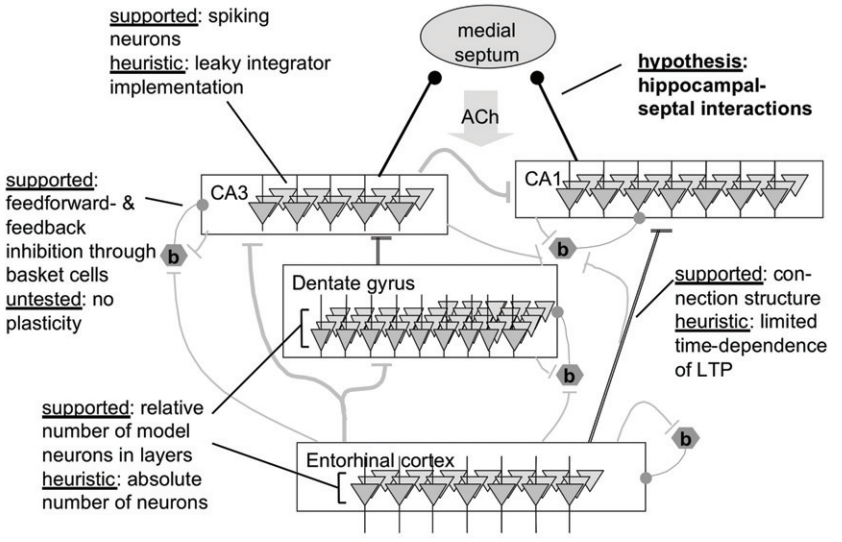


Figure 2. Examples of the different kinds of assumptions underpinning computational models. The neural network model of the hippocampus (also shown in Figure 1c) reproduces many features of hippocampal anatomy, such as the relative numbers of principal neurons in parts of intrahippocampal connections (*supported assumptions*). Other features of the model are clearly counterfactual. An example of such *heuristic assumptions* is that LTP occurs in the model only if postsynaptic spiking follows a few milliseconds after presynaptic firing, while some LTP occurs even for lags of 30–50 milliseconds. An example of an *untested assumption*, which could be true, is that there is no substantial plasticity in the connections to and from basket cells on the time scale of the model. Some untested assumptions concerned inhibitory interactions between the hippocampus and the medial septum. These assumptions, the *hypotheses*, were the focus of the paper in which the model was presented model (Meeter et al., 2004).

We will use the analysis above to discuss three sources of support for models: *a priori* plausibility, fitting data, and making predictions that turn out to be correct. These sources are not specific to computational work; verbal theories are also evaluated on how they explain data, on the fate of their predictions, and on how plausible they are. Presenting a theory in the form of a computational model has advantages over verbal theorizing, in that it makes hidden assumptions explicit, and makes it easier to detect fudges, ad hoc assumptions, and inconsistencies between the assumptions underlying the model. This explicitness makes it relatively easy to see what the model predicts when it is applied in new ways (in verbal theorizing, what counts as a prediction of the model is sometimes more hotly debated than the predictions themselves).

2.1. A Priori *Plausibility*

Outsiders would rather believe a model with assumptions that they know to be true, than one with assumptions that they have to accept without backup. The *a priori*

plausibility of a model is thus a function of having its assumptions supported by empirical evidence. Adding biological features does not in itself make a model more likely to be structurally adequate, however. If certain features are unimportant for the modeled behavior, to abstract away from them does not threaten structural accuracy. It is not the presence of biological features per se that renders a model plausible, but the presence of the relevant features as well as the absence of assumptions not in agreement with biology. This can be seen from an analysis of the process of model behavior derivation. In a technical sense, producing model behavior to fit data is similar to deriving a theorem from a set of postulates. The derived behavior allows the set of assumptions to be tested against data. By what is known as the Duhem–Quine thesis (Quine, 1951), however, a single hypothesis or set of hypotheses can never be tested in isolation; it is always tested together with the unavoidable background of theory and auxiliary hypotheses: both are necessary for the derivation of a concrete prediction. In the case of a model, these include the heuristic assumptions that are necessary for concrete model behavior, but are in all likelihood counterfactual. If it turns out that only the specific heuristic assumptions chosen allow the model to produce its target behavior, then the model is not structurally accurate (it does not produce the behavior with the same mechanisms as the brain). Deriving predictions from a model thus involves what one could call a meta-assumption of no commission: that the heuristic assumptions made are not critical for the model behavior used to derive a prediction. If this meta-assumption is true, many different sets of heuristic assumptions would, if combined with the evidence-based assumptions and hypotheses, lead to the same model behavior.

There are, of course, also sins of omission. Abstracting away from biological features may also ruin structural accuracy, if the elements left out in the abstraction turn out to be causally relevant for the modeled data. If a model reproduces target data, however, it can be concluded that the model contains at least enough features to produce those data. This is not to say that all relevant assumptions are true: the model may produce the behavior thanks to its heuristic assumptions. If a model fits data, sins of omission can thus only be present in the presence of sins of commission: if the right feature is not in the model, it takes a wrong feature to make up for it and still produce the right behavior. There is thus an asymmetry, where sins of commission have worse consequences for structural accuracy than sins of omission. *A priori* plausibility therefore hinges strongly on the absence of unsupported features, not the presence of supported ones.

2.2. Matching Data

Reproduction by the model of empirical data is a traditional standard for evaluating a model. It is referred to as the extent to which the behavior of the model 'matches' the to-be modeled data sets (Webb, 2001), or how well it 'fits' the data, or is 'descriptively adequate' (Chomsky, 1965; Jacobs & Grainger, 1994). How well a model fits data is the domain of a vast technical literature, which we will not attempt to summarize here (see Pitt, Myung, & Zhang, 2002; Zucchini, 2000). One important

aspect to mention, however, is that match is only impressive in an inflexible model (Roberts & Pashler, 2000). Flexibility refers to the amount of possible data patterns that a model could fit. If it could fit any conceivable data pattern, it is not a virtue that it can fit the true data.

In simulation models, flexibility is often hard to gauge. Traditionally, it has been estimated as the number of free parameters in the model, although it is now well established that this is a not a foolproof measure of how many data patterns a model can fit (Pitt et al., 2002). In psychobiological models it is usually not clear what counts as a free parameter and what not. For example, free parameters are often hidden in the translation of model behavior to target behavior. Consider time: much data have the form of a time duration (e.g., reaction times, interspike intervals) or is set in time (e.g., a learning curve). In a computational model processes take up a number of cycles or events, but the correspondence between such 'model time' and real time is usually addressed only in the scaling of graphs that show the fit between model and target behavior. Model time is fixed only in some low level models, where it is bound by neurophysiology (e.g., Hodgkin & Huxley, 1952).

Even a model with no free parameters may have been tinkered with until it produced the required behavior. That is, extra flexibility may be hidden in the structure of the model, although it is impossible to quantify how much. It is clear, however, that a sparse model, one with few assumptions, cannot have been tinkered with as extensively as a complex model. Moreover, assumptions that are bound to evidence also allow limited tinkering. We therefore suggest that the number of assumptions not bound by evidence (heuristic and untested assumptions, the latter including free parameter values) may be a rough measure of the flexibility of the model. This suggests that sparseness, as a proxy for inflexibility, is an essential asset for a model that is judged on its ability to match data.

What level of sparseness is required may depend on the number of data patterns explained by the model. Free parameters and heuristic assumptions might be forgivable in a model that explains a large number of independent data patterns. An example is a model of the hippocampus and adjacent cortical areas (Norman & O'Reilly, 2003). Although the model is not very sparse, this is made up by the fact that the model explained many separate findings in the memory literature.

2.3. Deriving Predictions

The strongest support a model or theory can receive is that daring predictions it makes are shown to be correct (Popper, 1934; Roberts & Pashler, 2000). The word 'prediction' has been subject to some inflation: many modelers call any model behavior a prediction, even when the data they are fitting have already been around for a long time (Roberts & Pashler, 2000). Here, we mean by a prediction something derived from the model behavior of which the modeler did not know whether it was the case. More suspicious minds may speak of predictions only if nobody knew, at the time that the model was presented, whether the prediction was true or not.

Computational models of the brain are seldom presented only after daring predictions have been proven to be correct (exceptions are Dehaene, Sergent, & Changeux, 2003; Rokers, Mercado, Allen, Myers, & Gluck, 2002). Many papers presenting a new model, however, close with a long list of predictions. These are then left open for experimentalists to confirm or refute (e.g., Bogacz et al., 2001; Hilgetag, 2000; Meeter & Murre, 2004; Raffone & Wolters, 2002; Rolls & Deco, 2002). Sometimes presentation of the model is followed by papers proving daring predictions correct (e.g., Frank, 2005; Frank, Seeberger, & O'Reilly, 2004, where, publication lags hide the earlier acceptance of the modeling paper). Although there are examples of models that made daring predictions and were supported through them, there are unfortunately also examples of predictions along the line of "more errors will be made in the difficult that in the easy condition." Of course, this is not the kind of model support we have in mind.

The likelihood that a model makes truly risky predictions is greatest if not only the modeler, but also others can derive predictions from the model. Whether this is possible is a function of how tightly the model is bound to its original domain. Modelers present their model as applied to one or more domain, such as a set of psycholinguistic tasks, a particular type of neuronal responses, or activity levels in a brain region generated by certain tasks. A good model is not bound to this 'native' domain, but is able to generate predictions outside it.

Jacobs and Grainger (1994) distinguish two ways in which such extension of the model can occur. Horizontal generality refers to the ability of models to be applied to new tasks, behavioral measures, or circumstances, vertical generality to the ability of a model to account for the behavior of the modeled system at different scales—different temporal scales or at different levels. For example, if a psychobiological model of memory links certain processes to particular brain regions it can generate novel predictions on the outcome of brain lesions (Gluck & Myers, 1993; Norman & O'Reilly, 2003). This will of course only work if the modeler is specific about what part of reality the whole model refers to. Else, failure of a model prediction can be disavowed on the basis that the model was misunderstood and did not really make the prediction that was tested. Such discussions do, in fact, occur in the literature.

3. Modeling: the Good, the Bad, and the Ugly

To summarize, models can be supported by *a priori* plausibility, their match to empirical data, or the generation of daring predictions.

- For *a priori* plausibility, as few assumptions as possible should lack support of empirical evidence.
- For a model to be supported by its match of the data, it should be sparse, making it inflexible relative to the quantity of data fitted. This may be the case because the model either contains few assumptions, or because all or many assumptions are evidence-based (i.e., bound by biology).

- For the model to be testable via the derivation of predictions, it should be vertically and horizontally general. For that, the ontology of the model should be clear.

From this analysis, two strategies for producing convincing models immediately become apparent. The first would be to build a sparse, inflexible model that can be genuinely tested against data using techniques that punish the model for flexibility (Pitt et al., 2002; Zucchini, 2000). The second would be to build a more complex model, but to bind as many assumptions as possible to biological evidence.

3.1. Good Modeling: Sparse Models

Sparse, inflexible models abound in the mathematical model literature (Cherniak, Changizi, & Kang, 1999; Chessa & Murre, 2002; Raaijmakers & Shiffrin, 1981; Shiffrin & Steyvers, 1997), for example, in low-level models of neurons (Hodgkin & Huxley, 1952; Kistler, Gerstner, & van Hemmen, 1997; Volny-Luraghi, Maex, Vosdagger, & De Schutter, 2002). There are also a few psychobiological models that explicitly strive for sparseness (Botvinick et al., 2001; Lengyel et al., 2005; Meeter, Myers, & Gluck, 2005). Botvinick et al. (2001), for example, added only one node to existing models, and with that small addition fitted several new data patterns.

The sparseness strategy is especially appropriate for models that are used to analyze data or as an existence proof. For example, some abstract models are used to derive forgetting functions from assumptions about the brain, These are then fitted on data, and used, for example, to analyze differences in forgetting between older and younger adults. Since forgetting data can be fitted by functions with just two free parameters, only very simple, sparse models can be used for such analysis. More in general, data analysis via a computational model is always only appropriate if the model is transparent and simple, so that the relation between the empirical data and the model outcome is clear. An existence proof is valid independent of its complexity or the support for its assumptions. Nevertheless, an existence proof is more compelling when it is transparent, and when few assumptions are necessary to produce the wished-for behavior. Transparency also makes it easier to link the model to brain and behavior.

3.2. Good Modeling: Binding to Biology

For many psychobiological models, the sparseness strategy is not plausible, however. They are too elaborate and flexible, precisely because their goal is broader than fitting only behavioral data or only brain data. The second strategy, which overlaps partly with the first, is to bind as many aspects of the model as possible to biology, by using evidence-based assumptions and by making the ontology of the model explicit. With this second strategy, modelers cannot rely on the match with data to support their model (although not reproducing existing data would of course falsify the model). Instead, modelers will have to rely on *a priori* plausibility, and on the ability of the

model to generate testable predictions. This second strategy thus allows the natural strength of psychobiological models, their vertical generality, to be played out by generating hypotheses at different levels.

In an extreme case, all assumptions are evidence-based. This would make the model's behavior plausible as a prediction independent of any fit of real data. This is no fata morgana. In fact, all models that purport to show computational constraints valid for the brain must follow this reasoning. If, for example, the memory capacity of the hippocampus can validly be inferred from the model of Treves and Rolls (1994), it must be the case that the model is wholly structurally accurate and contains all relevant assumptions. The validity of constraints derived by such a model critically depends on the meta-assumption of no commission and omission described above, that no heuristic assumption is critical for the behavior and no relevant assumption is left out. If these meta-assumptions are not true, the derived constraint may only apply to the model, not to the brain.

3.2.1. A requirement: ontological clarity

For the second strategy (binding assumptions to data) to be plausible, it must be clear what counts as an evidence-based assumption and what doesn't. Whether or not assumptions are supported can only be ascertained when it is clear what the model refers to in reality. If modelers are vague about what modules in their model stand for, it would be disingenuous to claim support from neuroanatomy for the architecture of the model. Similarly, if modelers are vague about the time scale of model events such as spiking, it would be disingenuous to claim that the inclusion of complex spike-time dependent learning rules makes the model plausible. For model assumptions to garner support, it is thus crucial that modelers are clear about the ontology of their model. This is even more true for generality: predictions can only be unambiguously derived from a model if it is clear to outsiders what it is a model of. If it is clear what a model refers to, on the other hand, it can be combined with other theories of the same substrate, and/or with any number of conjunctive hypotheses, leading to an in principle unbounded number of predictions that can be derived (Devitt, 1997). Whether or not these predictions can be tested with present techniques is of course another matter.

A useful notion here is that of the 'level of representation' of a model. Several lists for such levels have been proposed, for example one consisting of the neuron, network, map, system, and whole-brain level (Sejnowski & Churchland, 1993). In setting up such lists, there is usually no claim that each level has a distinct ontological status, or that phenomena at those levels are independent of one-another. Rather, each level is its own description of the same organ, the brain, that is the focus of research in neuroscience and psychology (Bakker & den Dulk, 1999). What counts as a level is often hard to decide, however. Size, the criterion used to discern levels in the list above, does not readily create levels of entities that interact with one another (e.g., Bechtel, 2007). Another approach would be to construct lists of epistemological levels, in which each level is addressed by different epistemological techniques (typically a discipline, as in the 'biological' and 'psychological' levels). These are then

sometimes held to be independent of one another (Putnam, 1973). Such a carving-up of levels does little good in the brain sciences, however, where interdisciplinary research is common. Such research would presumably either fall in between levels or create infinite new shadings of levels.

In the brain sciences, ontological and epistemological considerations can reinforce one another in deciding on levels. First, by giving a model a certain ontology, a modeler is also declaring the model vulnerable to testing with certain techniques. One cannot, for example, claim to model a brain part and then disavow predictions about what happens when that brain part is lesioned. Second, epistemological considerations can help define levels in an ontologically motivated list. Table 1 gives an example of a list of ontological levels driven by epistemological considerations. For each level, it specifies the entities modeled, but also the kinds of data that are addressed at that level. It identifies a behavioral, neuropsychological, circuit, neurophysiological and a biophysical level. The brain part sets a level (the neuropsychological level), and not for example the cortical column. This is because the brain part is the level at which a set of scientific methods applies (i.e., those of neuropsychology, animal lesion studies, and functional imaging), and this is not the case for the cortical column.[1]

In determining at what level a model is, one option is to look only at elemental units in the model that are said to correspond to something in reality. The level of these elements can then be given as the level of the model (Haefner, 1996). We propose a broader approach, in which every level at which the model purports to model something is a level of the model. If a model, for example, includes model neurons in a model hippocampus and a model hypothalamus, it would be at the neurophysiological level because it models neurons, and also at the circuit level because it models the physiology of brain parts. If it would also model some behavioral task, it would additionally be at the behavioral and neuropsychological level (as the model then also must predict what would happen to behavior if one of the two modeled structures were lesioned). A psychobiological model is one that is situated at both the behavioral level, and at least one level below it. The more levels of representation a model has, the more data can refute the model.

Psychobiological modelers are often not very clear at which levels they see their model—whether or not, for example, their units can be seen as a sample of real neurons, or whether or not fMRI can falsify their model. The imprecision may concern the structures modeled (e.g., stating that a model stands for the 'Medial Temporal Lobe' or the 'visual system' without specifying what belong to these regions or systems). Moreover, heuristic assumptions can invalidate the comparison of the behavior of model parts with activity in the brain regions modeled. For example, if model nodes in a neural network have both positive and negative activations (as in the Hopfield network), it becomes difficult to compare node activation with neural firing in the target structure (Hasselmo, 1995b). Such heuristic assumptions may thus make derivation of predictions at a low level impossible. Imprecision can also be about the organism modeled (e.g., a model matches only human behavior, but the underpinning of its assumptions come from animal physiology). Although it can be

Table 1. Example of a list of modeling levels, with for each level the substrates modelled, the data that are addressed at the level, and some examples of published models at that level.

Level	What is modeled?	Data addressed	Examples
Behavioral	Behavior whole organism	Behavioral	Shiffrin & Steyvers (1997); Bundesen (1990)
Neuropsychological	Functions of identifiable brain parts	Lesion studies, fMRI findings	Gluck & Myers (1993); Mozer, (1999); Mozer et al. (1997); Norman & O'Reilly (2003)
Circuit	Processes in identifiable brain parts	Gross brain anatomy, cFOS, spike counts in electrophysiological recordings	Deco & Rolls (2002); Fincham et al. (2002); Grossberg (2001); Roelfsema et al. (2002); Xing & Andersen (2000); Li (2003); Frank (2005)
Neurophysiological	Neurons at specified location, and dynamics of those neurons	Fine brain anatomy, electrophysiology, neuropharmacology	Brunei & Wang (2001); Dehaene et al. (2003); Hasselmo et al. (2002); Szabo et al. (2004)
Biophysiological	Components of real, identifiable neurons	Physiological data at subcellular level	Hodgkin & Huxley (1952); Volny-Luraghi et al. (2002); Polsky et al. (2004)

claimed that a model is true of the target structure in several related species, such generality cannot be taken for granted or deduced from similarities in behavior (see Treves & Samengo, 2002 for a counter example). Lastly, the imprecision may concern aspects of the events modeled, for example if the time scale of the simulation is unclear.

An interesting example of prediction outside of the domain of the model, despite ontological vagueness of three kinds, is offered by a model of human recognition memory (Norman & O'Reilly, 2003). This model consists of a fairly worked-out hippocampal model, and a simple, quite abstract neocortical module that is identified with the 'medial temporal lobe neocortex'. In this last module, patterns are stored in a set of weights between an input and an output layer. If a pattern is presented for a second time, it leads to higher activity in the 'winners' in the output layer of the neocortical region, which is interpreted as a familiarity signal.

In their presentation of this model, Norman and O'Reilly note an apparent contradiction between this mechanism and data from cell recordings from macaque perirhinal cortex. Xiang and Brown (1998) showed decreased firing for familiar patterns in perirhinal neurons. The explanation offered by Norman and O'Reilly for this contradiction is less remarkable than the fact that they felt compelled to comment on it. Their model is targeted at human behavioral data, is rather abstract, and does not include activity measures on an explicit time scale. The fact that Norman and O'Reilly see monkey electrophysiology as relevant for their model implies that they see their model as situated on at least the circuit level in Table 1, and describing both the human and the macaque perirhinal cortex. Given that they derived successful predictions at the behavioral and neuropsychological levels, their model enjoys considerable vertical generality.

3.3. Modeling, the Ugly

The analysis of model support does not only show ways in which computational modeling can be good, but also its corollary: ways in which modeling can be bad. A complex model that explains little data is clearly not one that we will learn much from. But is such a model automatically so bad that it does not deserve to be put out on the intellectual market place? Modelers may sometimes not be able to construct a model supported by and explaining lots of data. To base one's model on and use one's model to explain empirical evidence, that evidence must also be there. At low levels, data are often available for this purpose (e.g., many parameter values in the Hodgkin-Huxley formalism can be extracted from experimental data). At higher levels, however, there is no abundance of restricting data. Both neuropsychological and imaging data are still relatively sparse, imprecise, often disputed and not well understood. At these levels, modeling is a relatively free, unconstrained activity (Murre, 2002). The staggering variety in published computational models at these levels attests to this freedom.

Should modeling in these areas then not just wait until more data are available? We would argue that there is a case for models that are not bad, but 'ugly': psychobiological models that are not tied very neatly to biology, nor are very sparse,

nor explain that much data, but that are the first in their area and a stab in the right direction on all three fronts. Such models can function as a seed for further efforts. For example, in the area of the consolidation of long-term memory, the first computational model was that of Alvarez and Squire (1994). It was not very close to biology, nor very sparse, and explained only a single data pattern (the gradient often seen in retrograde amnesia after damage to the medial temporal lobe). Nevertheless, this model made a hitherto vague verbal theory—that memories are consolidated from the hippocampus to the neocortex—explicit, and provided a start for later models, which explained more data and were often closer to biology, to build on (McClelland et al., 1995; Meeter & Murre, 2005) or argue against (Nadel et al., 2000).

4. If All Else Fails: Model Hierarchies

For modelers that attempt to explain behavior from brain mechanisms, it may be tempting to retreat to lower levels where evidence is more abundant. However, incorporating low-level data into a model that must also provide explanations for behavior may make the model unwieldy. Consider what would have happened if Norman and O'Reilly (2003) had set out to change their memory model to account for the electrophysiological data provided by Xiang and Brown (1998). It is very unlikely that assumptions needed to explain neurophysiology would be of much help in explaining the data targeted by Norman and O'Reilly (2003), such as mirror effects in recognition memory. They would have needed one set of assumptions to explain the phenomena at the neuronal level, and then another set to explain behavioral effects. Possibly, yet a third set of assumptions would have been needed to bridge the intermediate levels. Because many of these assumptions would have lacked supporting data, this would have made their model top-heavy with heuristic assumptions. In such cases, when plausible biological underpinnings are not available or their inclusion requires the addition of many untested and heuristic assumptions, it may be preferable to disregard some facts in order to construct a simpler model (following the first strategy). This was done, in fact, by Norman and O'Reilly (2003).

There is an alternative to falling back on an abstract model remote from biological reality, or to building a complex, unconstrained biological theory that needs many untested and heuristic assumptions to work. One may tear the complex, unconstrained model apart into several simpler models at different levels. In this way, a hierarchy of models is created, where each model is a more or less detailed elaboration of the same idea. Models in the hierarchy could then constrain one-another and impose restrictions where behavior and neurobiology pose too few. This is the essence of what we call the model hierarchy approach.

4.1 A Model Hierarchy

The central idea of model hierarchies is that a theory is not implemented in a single model, but in several models at varying levels of representation. Lower-level models in this family are concretizations of higher-level models, and higher-level models are

abstractions, simplifications of the lower-level models. An example of such relations, outside of computational modeling, is the classical theory of optics, which describes light propagation in terms of rays, and the theory of electromagnetic radiation, describing propagation in terms of electromagnetic waves. The wave theory is more correct, but classical optics is widely regarded as useful at a macro level, explaining, for instance, how light is reflected or bent at the interface between two dissimilar media. There is thus added value to having two hierarchically organized versions of the same theory. Another example, to which we will return below, is the relation between the biochemistry of DNA and behavioral genetics.

The relations between models at different levels in the proposed hierarchies are similar to those between different levels in the hierarchy of Marr (1982), i.e., between the implementation and algorithmic levels, or between the algorithmic and competence levels. A lower-level model can, for example, be an algorithmic implementation of a higher-level model, with the higher-level model specifying the competences of the lower-level model. In Marr's view the levels are independent, whereas relations between models in the hierarchy imply that no assumption in one model may be in contradiction with assumptions or behavior of the others. Ideally, all higher-level models would in principle be translatable into the framework of lower level models, without loss of function. For example, if a modeler had unlimited time and unlimited computational resources, all elements in his or her behavioral level model could be replaced by the most low-level biologically-grounded elements simulated at a millisecond scale, and the model would still be able to simulate behavioral phenomena occurring at a scale of minutes and hours. Although this is usually an unattainable ideal, given that any model needs heuristic assumptions to work, it is one with force. It excludes the use of some heuristic assumptions that preclude translation of the model into a lower-level one, such as the use of negative activations, negative weights, and error-correcting learning.

The idea of replacing all elements of a higher-level model by elements of a lower-level model was used by Putnam (1973) to support the independence of levels, a claim also defended by Marr (1982). Putnam's example was a wooden board with two holes. The task was to explain why a peg goes through one and not the other. The correct explanation, he said, was that one hole was shaped right and one was not. Trying to improve on this explanation by going down, through the molecules in the wood, to quantum theory would, even if feasible, only be silly. In similar fashion, explanations for psychological phenomena should reside at the psychological level, not try to go below it. The reductionism that Putnam was arguing against (trying to reduce psychology to natural sciences) is now daily practice. Nevertheless, part of his argumentation, that trying to explain a high level phenomenon from a low-level phenomenon leads to unwieldy complex accounts, is exactly the problem that led us to suggest model hierarchies.

What advantages do model hierarchies have that make them better than Putnam's solution of rejecting reductionism and sticking to one level? In many cases, they come down to that a wooden peg with holes is a particularly bad metaphor for the scientific issues in psychology. Serious questions in psychology often do not have obvious

answers at the psychological level.[2] If the answer to many of these questions involves the brain, as many psychologists assume, a psychobiological model that connects brain and behavior may be called for. In case where the data gap occurs, there are too few biological and behavioral constraints to develop either abstract, testable models or elaborate theories. In such cases, a modeling hierarchy may allow the development of a coherent framework that forms a viable theory. By splitting up the model in a model hierarchy, each model in that hierarchy can be simple and transparent, and data at all levels can be incorporated in or accounted for by the model. Behavior of the lower level model may feature as an untested assumption in the higher-level model. Without the lower level, this would just be one of possibly many assumptions that an outsider just has to accept or reject. With the lower-level model, the plausibility of the untested assumption can be assessed. Moreover, support for the lower level model would strengthen the higher-level model, and rejection of the lower-level model would mean at least a reconsidering of the higher-level model.

Above, it was mentioned that replacement by elements of high-level model structures by lower-level structures is usually an unattainable ideal: Attempting to translate the higher-level model into the lower level model will usually uncover small inconsistencies between them.[3] To go back to an example outside of the modeling literature, the methods of behavioral genetics assume that there is no change in genes when they are passed from parent to offspring. Meanwhile the biochemistry of DNA has enlightened us on all sorts of changes that occur in DNA as it is passed on. The claim that behavioral genetics and the biochemistry nevertheless stand in a hierarchical relation entails the further claims that the inconsistencies are insubstantial, and that when they arise the lower-level model is always right. Both are clearly beliefs of behavioral geneticists, who will readily admit that biochemistry has the last word on DNA, but believe that mutations are sufficiently rare not to invalidate their methods (Martin, Boomsma, & Machin, 1997; Plomin, DeFries, McClearn, & McGuffin, 2001).

So what, in the end, makes two models stand in a hierarchical relationship to one-another? Both models must be of the same substrate, with the lower-level model either modeling exactly the same part of the brain as the higher-level model, or an identifiable part of the higher-level model. Second, assumptions of the one model may not be in substantial contradiction with the other model. Third, the lower-level model must explain untested assumptions of the higher-level model. Whether this is the case is in first instance a matter of the modeler claiming that this is so. In second instance, his or her claims are empirical (they relate to the factual relations between models), and can be criticized by other scientists.

4.2. Examples of Hierarchies

Several examples of such hierarchies have already been presented in the literature. In the literature on retrograde amnesia (remote memory loss), biological information only weakly constrains modeling, and the target behavioral data consist of a qualitative pattern of just two curves (the decreasing forgetting curve and

increasing Ribot curve of graded retrograde amnesia). Here, accounting for empirical data while at the same time presenting a more basic model has been accomplished with two hierarchies of models. McClelland et al. (1995) presented a backpropagation network as an existence proof for their central mechanism, interleaved learning.[4] With a mathematical abstraction of the same process, they fitted several retrograde amnesia curves to show how their process would apply in the real world. Meeter and Murre (2004; 2005; Murre, 1996) presented another neural network model of amnesia, TraceLink. The assumptions underlying TraceLink were also included in a concise, mathematical model of learning and forgetting (Murre et al., 2007). By fitting this model to curves from the amnesia literature, the theory is explicated on the behavioral level, leading to, for example, estimates of the time course of consolidation.

Moreover, TraceLink contains two untested assumptions about learning that were, at the level of the TraceLink model, not testable (the neuropsychological level in Table 1). It assumes that hippocampal codes are independent of the neocortical ones (i.e., if two patterns have similar neocortical representations, they will nevertheless have orthogonal hippocampal ones), and that modulation of learning produces high learning rates for novel patterns in the hippocampus, and low learning rates for old patterns. Both assumptions were fleshed out in separate, lower-level models, the first in a model of the parahippocampal gyrus (Talamini, Meeter, Murre, Elvevåg, & Goldberg, 2005), the second in a model of cholinergic modulation of the hippocampus (Meeter, Talamini, & Murre, 2004) inspired by ideas of Hasselmo (e.g., 1995a). In both cases, the lower level model makes explicit what needs to be the case for the higher-level TraceLink model to be true. Moreover, they turn an untested—and untestable—assumption of TraceLink into testable behavior of a lower level (see Figure 3).

With these two hierarchies, not many data were available to constrain modeling. In more developed areas, model hierarchies can also play a useful role by allowing a modeler to combine the virtues of abstract, high level models with those of an inclusive, vertically general model-as-theory, much like the approach taken in other domains of science, such as optics or genetics. An example is a mean-field derivation of firing rates (Amit & Brunel, 1997; Brunel & Wang, 2001), which determines the average discharge rate of populations of spiking neurons in a network, simplifying its calculations. Another example is the relation between detailed modeling of spike generation allowed by the very complex Hodgkin–Huxley formalism (Hodgkin & Huxley, 1952), and the much simpler formalism of spike response models. Both model currents and spike formation in neurons, but the Hodgkin–Huxley formalism does this in more detail and at a smaller time scale than the spike response model. By showing that these models produce approximately the same behavior, Kistler et al. (1997) showed that the equations of their spike response model approximate the behavior of the Hodgkin–Huxley model neuron. The authors could then use the validity of the Hodgkin–Huxley model to argue for their own equations, changing them from untested assumptions to supported assumptions. Their work allows modelers to use these neuron derivations, enjoy their lower computational load and

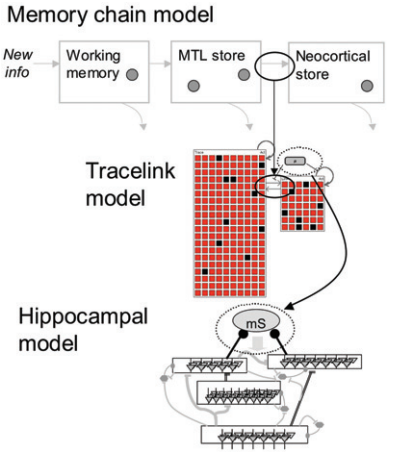


Figure 3. Example of a model hierarchy of three models of hippocampal involvement in memory (also shown in Figure 1). Features that are untested assumptions at higher levels are fleshed out at lower levels. The two examples shown are the assumption that memories are copied from an MTL store to a neocortical store, fleshed out in the Tracelink model, and the assumption of a modulatory system controlling plasticity in the hippocampus, fleshed out in the more detailed model of the hippocampus and its interactions with the medial septum.

reduced set of parameters, and still have the same structural accuracy as when they had used the original Hodgkin–Huxley model neurons. In this way, the virtues of an explicit, low-level theory and a sparse, simple model can be combined.

5. Conclusion

Computational models of the brain hold great promise as a bridge between biology and psychology. Many models that attempt to bridge the gap, those that we referred to as psychobiological, have already been proposed. If these models are to fulfill this promise, however, they will have to be structurally accurate. If model layers stand for cortical regions without there being a clear relation between layer behavior and that region, what can we learn from such a model? That the model can generate the target behavior while the brain works in some mysterious other way is surely not a worthwhile lesson.

Structural accuracy is the goal of the model, not something that can be ascertained beforehand. Three ways to support structural accuracy were discussed. A model may be *a priori* plausible because of its use of evidence-based assumptions. It may also be supported by its fit of existing data provided the fits are not caused by an excess of flexibility in the model. Finally, it may be supported by the confirmation of risky predictions generated by the model. For all three sources of support, it was argued

that models need to be bound to biology. As 'biological plausibility' is more often claimed than substantiated (Jacobs & Grainger, 1994), such a claim is not enough. Instead, modelers must be explicit about the ontology of their model, and use assumptions supported by data.

In some cases there may not be data available to support assumptions. We have suggested a strategy for avoiding the construction of unconstrained cathedrals of speculation. In this strategy, a hierarchy of models is developed that all share the same assumptions, and higher-level models can in principle be translated into the formalism of the lower-level models. In a sense, such a hierarchy incorporates the goals of classical reductionism, where constructs of a higher level are identified with entities on lower levels. These hierarchies of models allow the development of theory without unrestricted dabbling in speculative theorizing. Moreover, they may allow a combining of the virtues of sparse, testable models and vertically general, biology-rich models. Modeling the same substrate at different levels of representation, as proposed here, may thus have benefits that exceed the merits of each model in the hierarchy on its own.

Acknowledgements

This research was supported by a VENI grant to the first author and a PIONIER grant to the third author, both from the Netherlands Society for Scientific Research (NWO).

Notes

[1] There are computational models that operate at the level of cortical columns, but these have generally failed to get traction; perhaps because of a mismatch with available empirical techniques.

[2] Even when they do the answers are usually still in need of an explanation. For example, memory decay may be part of the psychological answer to the psychological question of why we forget, it is still an interesting question how that decay occurs. To that question a psychological answer is unlikely.

[3] This is generally the case for theories that describe the same phenomena at two levels, as the debate on reductionism has shown (e.g., Schaffner, 1967; Sklar, 1967).

[4] Interleaved learning refers to mixing learning trials for new patterns with repetition trials for old, already stored patterns.

References

Alvarez, R., & Squire, L. R. (1994). Memory consolidation and the medial temporal lobe: a simple network model. *Proceedings of National Academy of Sciences (USA)*, *91*, 7041–7045.

Amit, D. J., & Brunel, N. (1997). Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex. *Cerebral Cortex*, *7*, 237–252.

Anderson, J. R. (2002). Spanning seven orders of magnitude: A challenge for cognitive modeling. *Cognitive Science*, *26*, 85–112.

Bakker, B., & den Dulk, P. (1999). Causal relationships and relationships between levels: The modes of description perspective. In M. Hahn & S. C. Stoness (Eds.), Proceedings of the Twenty-First Annual Conference of the Cognitive Science Society (pp. 43–48).

Bechtel, W. (2007). Reducing psychology while maintaining its autonomy via mechanistic explanations. In M. Schouten & H. Looren de Jong (Eds.), *The matter of the mind: Philosophical essays on psychology, neuroscience and reduction* (pp. 172–198). Oxford: Blackwell Publishing.

Bogacz, R., Brown, M. W., & Giraud-Carrier, C. (2001). Model of familiarity discrimination in the perirhinal cortex. *Journal of Computational Neuroscience*, *10*, 5–23.

Botvinick, M. W., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, *108*, 624–652.

Brunel, N., & Wang, X. J. (2001). Effects of neuromodulation in a cortical network model of object working memory dominated by recurrent inhibition. *Journal of Computational Neuroscience*, *11*, 63–85.

Bundesen, C. (1990). A theory of visual attention. *Psychological Review*, *97*, 523–547.

Cherniak, C., Changizi, M., & Kang, D. W. (1999). Large-scale optimization of neuron arbors. *Physical Review*, *59*, 6001–6009.

Chessa, A. G., & Murre, J. M. J. (2002). A model of learning and forgetting, I: The forgetting curve. Amsterdam: NeuroMod Technical Report 02-01.

Chomsky, N. (1965). *Aspects of a theory of syntax*. Cambridge: MIT Press.

Deco, G., & Rolls, E. T. (2002). Object-based visual neglect: A computational hypothesis. *European Journal of Neuroscience*, *16*, 1994–2000.

Dehaene, S., Sergent, C., & Changeux, J. P. (2003). A neuronal network model linking subjective reports and objective physiological data during conscious perception. *Proceedings of the National Academy of Sciences USA*, *100*, 8520–8525.

Devitt, M. (1997). *Realism and truth* (2nd ed.). Princeton: Princeton U.P.

Fincham, J. M., Carter, C. S., van Veen, V., Stenger, V. A., & Anderson, J. R. (2002). Neural mechanisms of planning: A computational analysis using event-related fMR. *Proceedings of the National Academy of Sciences USA*, *99*, 2246–3351.

Frank, M. J. (2005). Dynamic dopamine modulation in the basal ganglia: A neurocomputational account of cognitive deficits in medicated and non-medicated Parkinsonism. *Journal of Cognitive Neuroscience*, *17*, 51–72.

Frank, M. J., Seeberger, L., & O'Reilly, R. C. (2004). By carrot or by stick: Cognitive reinforcement learning in Parkinsonism. *Science*, *306*, 1940–1943.

Frankland, P. W., O'Brien, C., Ohno, M., Kirkwood, A., & Silva, A. J. (2001). a-CaMKII-dependent plasticity in the cortex is required for permanent memory. *Nature*, *411*, 309–313.

Gluck, M. A., & Myers, C. (1993). Hippocampal mediation of stimulus representation: A computational theory. *Hippocampus*, *3*, 491–516.

Grossberg, S. (2001). Linking laminar circuits of visual cortex to visual perception: Development, grouping, and attention., 25, 613–526. *Neuroscience and Biobehavioral Reviews*, *25*, 513–526.

Haefner, J. W. (1996). *Modeling biological systems: Principles and applications*. New York: Chapman & Hall.

Hasselmo, M. E. (1995a). Neuromodulation and cortical function: Modeling the physiological basis of behavior. *Behavioural Brain Research*, *67*, 1–27.

Hasselmo, M. E. (1995b). Physiological constraints on models of behavior. In L. F. Niklasson & M. B. Bodén (Eds.), Current trends in connectionism: Proceedings of the Swedish Conference on Connectionism—1995 (pp. 15–32). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hasselmo, M. E., Bodelón, C., & Wyble, B. P. (2002). A proposed function for hippocampal theta rhythm: Separate phases of encoding and retrieval enhance reversal of prior learning. *Neural Computation*, *14*, 793–817.

Hempel, C. G. (1965). *Aspects of scientific explanation: And other essays in the philosophy of science*. New York: Free Press.

Hilgetag, C. C. (2000). Spatial neglect and paradoxical lesion effects in the cat - A model based on midbrain connectivity. *Neurocomputing*, *32–33*, 793–799.

Hodgkin, A. L., & Huxley, A. F. (1952). A quantitative description of ion currents and its application to conduction and excitation in nerve membranes. *Journal of Physiology (London)*, *117*, 500–544.

Jacobs, A. M., & Grainger, J. (1994). Models of visual word recognition-sampling the state of the art. *Journal of Experimental Psychology: Human Perception and Performance*, *20*, 1311–1334.

Jensen, O., Idiart, M. A., & Lisman, J. E. (1996). Physiologically Realistic formation of autoassociative memory in networks with theta/gamma oscillations: The role of fast NMDA channels. *Learning & Memory*, *3*, 243–256.

Kistler, W., Gerstner, W., & van Hemmen, J. L. (1997). Reduction of Hodgkin Huxley equations to a single-variable threshold model. *Neural Computation*, *9*, 1015–1045.

Lengyel, M., Kwag, J., Paulsen, O., & Dayan, P. (2005). Matching storage and recall: Hippocampal spike-timing dependent plasticity and phase response curves. *Nature Neuroscience*, *8*, 1677–1683.

Li, Z. (2003). V1 mechanisms and some figure-ground and border effects. *Journal of Physiology – Paris*, *97*, 503–515.

Marr, D. (1982). *Vision*. San Franscisco: W.H. Freeman.

Martin, N. G., Boomsma, D. I., & Machin, G. A. (1997). A twin-pronged attack on complex traits. *Nature Genetics*, *17*, 387–392.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*, 419–457.

Meeter, M., & Murre, J. M. J. (2004). Simulating episodic memory deficits in semantic dementia with the TraceLink model. *Memory*, *12*, 272–287.

Meeter, M., & Murre, J. M. J. (2005). TraceLink: A connectionist model of consolidation and amnesia. *Cognitive Neuropsychology*, *22*, 559–588.

Meeter, M., Murre, J. M. J., & Janssen, S. M. J. (2005). Remembering the news: Modeling retention data from a study with 7500 participants. *Memory & Cognition*, *33*, 793–810.

Meeter, M., Myers, C. E., & Gluck, M. A. (2005). Integrating incremental learning and episodic memory models of the hippocampal region. *Psychological Review*, *112*, 560–585.

Meeter, M., Talamini, L. M., & Murre, J. M. J. (2004). Mode shifting between storage and recall based on novelty detection in oscillating hippocampal circuits. *Hippocampus*, *14*, 722–741.

Mozer, M. C. (1999). Explaining object-based deficits in unilateral neglect without object-based frames of reference. In J. A. Reggia, E. Ruppin, & D. Glanzman (Eds.), *Progress in Brain Research* (Vol. 121, pp. 99–119). Amsterdam: Elsevier Sciende.

Mozer, M. C., Halligan, P. W., & Marshall, J. C. (1997). The end of the line for a brain-damaged model of unilateral neglect. *Journal of Cognitive Neuroscience*, *9*, 171–190.

Murre, J. M. J. (1996). TraceLink: A model of amnesia and consolidation of memory. *Hippocampus*, *6*, 675–684.

Murre, J. M. J. (2002). Connectionist models of memory disorders. In A. D. Baddeley, B. Wilson, & M. D. Kopelman (Eds.), *Handbook of Memory Disorders* (2nd ed., pp. 101–122). New York: John Wiley.

Murre, J. M. J., Meeter, M., & Chessa, A. G. (2007). Modeling amnesia: connectionist and mathematical approaches. In M. J. Wenger & C. Schuster (Eds.), *Statistical and Process Models for Neuroscience and Aging* (pp. 119–162). Mahwah, NJ: Lawrence Erlbaum.

Murre, J. M. J., & Sturdy, D. P. F. (1995). The connectivity of the brain: Multi-level quantitative analysis. *Biological Cybernetics*, *73*, 529–545.

Nadel, L., Samsonovitch, A., Ryan, L., & Moscovitch, M. (2000). Multiple trace theory of human memory: Computational, neuroimaging and neuropsychological results. *Hippocampus*, *10*, 352–368.

Norman, K. A., & O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: A complementary learning systems approach. *Psychological Review*, *110*, 611–646.

Petrov, A. A., Dosher, B. A., & Lu, Z. L. (2005). The dynamics of perceptual learning: An incremental reweighting model. *Psychological Review*, *112*, 715–743.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, *109*, 472–491.

Plomin, R., DeFries, J. C., McClearn, G. E., & McGuffin, P. (2001). *Behavioral genetics* (4th ed.). New York: Worth.

Polsky, A., Mel, B. W., & Schiller, J. (2004). Computational subunits in thin dendrites of pyramidal cells. *Nature Neuroscience*, *7*, 621–627.

Popper, K. R. (1934). *Logik der Forschung: zur Erkenntnistheorie der modernen Naturwissenschaft*. Wien: Springer.

Putnam, H. (1973). Reductionism and the nature of psychology. *Cognition*, *2*, 131–146.

Quine, W. V. O. (1951). Two Dogmas of Empiricism. *Philosophical Review*, *60*, 20–43.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review*, *88*, 93–134.

Raffone, A., & Wolters, G. (2002). A cortical mechanism for binding in visual working memory. *Journal of Cognitive Neuroscience*, *13*, 766–785.

Rao, R. P. N., Zelinsky, G., Hayhoe, M., & Ballard, D. H. (2002). Eye movements in visual search. *Vision Research*, *42*, 1447–1463.

Ringo, J. L., Doty, R. W., Demeter, S., & Simard, P. Y. (1994). Time is of the essence: A conjecture that hemispheric specialization arises from interhemispheric conduction delay. *Cerebral Cortex*, *4*, 331–343.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*, 358–367.

Roelfsema, P. R., Lamme, V. A. F., Spekreijse, H., & Bosch, H. (2002). Figure-ground segregation in a recurrent network architecture. *Journal of Cognitive Neuroscience*, *14*, 525–537.

Rokers, B., Mercado, E., Allen, M. T., Myers, C. E., & Gluck, M. A. (2002). A connectionist model of septohippocampal dynamics during conditioning: Closing the loop. *Behavioral Neuroscience*, *116*, 48–62.

Rolls, E. T., & Deco, G. (2002). *Computational neuroscience of vision*. Oxford: Oxford University Press.

Rubin, D. C., & Wenzel, A. E. (1996). One hundred years of forgetting: A quantitative description of retention. *Psychological Review*, *103*, 734–760.

Schaffner, K. F. (1967). Approaches to reductionism. *Philosophy of Science*, *65*, 137–147.

Sejnowski, T. J., & Churchland, P. S. (1993). Brain and cognition. In M. I. Posner (Ed.), *Foundations of cognitive science* (pp. 301–356). Cambridge, MA: MIT Press.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. *Psychonomic Bulletin & Review*, *4*(2), 145–166.

Sklar, L. (1967). Types of inter-theoretic reduction. *British Journal for the Philosophy of Science*, *18*, 109–124.

Szabo, M., Almeida, R., Deco, G., & Stetter, M. (2004). Cooperation and biased competition model can explain attentional filtering in the prefrontal cortex. *European Journal of Neuroscience*, *19*, 1969–1977.

Talamini, L. M., Meeter, M., Murre, J. M. J., Elvevåg, B., & Goldberg, T. E. (2005). Integration of parallel input streams in parahippocampal model circuits; implications for schizophrenia. *Archives of General Psychiatry*, *62*, 485–493.

Treves, A., & Rolls, E. T. (1994). Computational analysis of the role of the hippocampus in memory. *Hippocampus*, *4*, 374–391.

Treves, A., & Samengo, I. (2002). Standing on the gateway to memory: Shouldn't we step in?. *Cognitive Neuropsychology*, *19*, 557–575.

Usher, M., & Niebur, E. (1996). Modeling the temporal dynamics of IT neurons in visual search: A mechanism for top-down selective attention. *Journal of Cognitive Neuroscience*, *8*, 311–327.

van der Velde, F., & de Kamps, M. (2001). From knowing what to knowing where: Modeling object-based attention with feedback disinhibition of activation. *Journal of Cognitive Neuroscience*, *13*(4), 479–491.

Volny-Luraghi, A., Maex, R., Vosdagger, B., & De Schutter, E. (2002). Peripheral stimuli excite coronal beams of Golgi cells in rat cerebellar cortex. *Neuroscience*, *113*, 363–373.

Webb, B. (2001). Can robots make good models of biological behaviour?. *Behavioral and Brain Sciences*, *24*(6), 1033–1094.

Xiang, J. Z., & Brown, M. W. (1998). Differential neuronal encoding of novelty, familiarity, and recency in regions of the anterior temporal lobe. *Neuropharmacology*, *37*, 657–676.

Xing, J., & Andersen, R. A. (2000). Models of the posterior parietal cortex which perform multimodal integration and represent space in several coordinate frames. *Journal of Cognitive Neuroscience*, *12*(4), 601–614.

Zucchini, W. (2000). An introduction to model selection. *Journal of Mathematical Psychology*, *44*, 41–61.